DOCUMENT RESUME

ED 294 898 TM 011 549

AUTHOR Shannon, Gregory A.
TITLE Usefulness of Score Interpretive Information for Examinees Who Fail Criterion-Referenced Tests.
SPONS AGENCY Educational Testing Service, Princeton, NJ. Office of Corporate Quality Assurance.
PUB DATE Apr 86
NOTE 93p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, April 1986); For a related document, see TM 011 550.
PUB TYPE Speeches/Conference Papers (150) -- Reports - Evaluative/Feasibility (142) -- Tests/Evaluation Instruments (160)

EDRS PRICE MF01/PC04 Plus Postage.
DESCRIPTORS *Criterion Referenced Tests; Failure; *Feedback; *Scoring; *Test Interpretation
IDENTIFIERS Educational Testing Service

ABSTRACT

The types of test score interpretive information considered useful to failing examinees were studied through interviews with Educational Testing Service (ETS) staff members. Research literature on interpretive information for failing test takers was reviewed, and procedures currently used at ETS were determined. Managers of 23 testing programs commented on score information currently provided. First among suggestions for improvement were tightening definitions of test content areas and developing content area scores consistent with total score results. Some examples of score interpretive information are included. Appendices include the project managers' interview questionnaire and a bibliography. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED294898

Usefulness of Score Interpretive Information for Examinees Who Fail Criterion-referenced Tests

Gregory A. Shannon

Educational Testing Service

Paper presented at the Annual Meeting of the
American Educational Research Association
San Francisco, California, April 1986

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

GREGORY A. SHANNON

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

This project was funded by the Office of Corporate Quality Assurance, Educational Testing Service.

TM 011 549

BEST COPY AVAILABLE

ACKNOWLEDGEMENTS

The author wishes to express his gratitude to Jerome R. Murphy, Director, Office of Corporate Quality Assurance, for his invaluable contributions as the major source of information about the audit process and the nature of the problem under investigation, and for his encouragement and support. Special thanks are offered to Alfred B. Carlson, Executive Director, COPA Statistical Area, for his sound advice throughout the duration of this project. Appreciation is also extended to Barbara Kirsh, Assistant Director, Office of Corporate Quality Assurance, for being instrumental in the selection of former auditors for the interviews, as well as for offering useful feedback at various stages of the project. The author also expresses his appreciation to Karin L. Abbey for her constructive editorial services, David C. Shlegle for technical support, and to the following individuals for their noteworthy contributions as interviewees and reviewers:

Interviewees

Robert F. Boldt
Thomas F. Donlon
Gary L. Marco
S. Donald Melville
Melinda A. Peabody
Rufus B. Smith, Jr.
Charles J. Teryek
Patricia L. Testa
Carolyn Y. Uhrich
Leo F. Walrad

Reviewers

Robert F. Boldt
Richard L. Burns
Daniel R. Eignor
Samuel A. Livingston
Gerald J. Melican
S. Donald Melville
Craig N. Mills
Melinda A. Peabody
Benjamin Shimberg
Rufus B. Smith, Jr.
Richard F. Thornton
Leo F. Walrad
Michael J. Zieky

SUMMARY

The study addresses ETS Score Interpretation Guideline 1, as it pertains to reporting score interpretative information to failing examinees. The major purpose of the study is to identify types of score interpretative information considered to be useful to failing examinees. The following information gathering activities are described:

1. Review of Research Literature

2. Survey of Reporting Procedures in Use at COPA

3. Interviews with COPA Testing Program and Operations Managers

4. Interviews with ETS Staff Familiar with Problems of Auditing Score Interpretation Guideline 1

Recommendations are offered to assist with the interpretation of the guideline and with the evaluation of the appropriateness of score interpretative information for failing examinees; an abbreviated list of recommendations is presented for program directors, test developers and measurement statisticians. Examples of exemplary score interpretative information are also provided.

Table of Contents

Page

I. OVERVIEW OF THE PROBLEM 1

Background 1
Score Interpretation Procedural Guideline 1 2
A Matter of Interpretation 2
Purpose and Scope of Study 4

II. DISTINCTION BETWEEN CRITERION-REFERENCED AND DIAGNOSTIC TESTS 6

III. REPORTING OF CRITERION-REFERENCED TEST RESULTS 10

IV. WHAT DIAGNOSTIC INFORMATION CAN BE PROVIDED BY CRITERION-REFERENCED TESTS 13

Percentage-Correct Scores 16

V. REPORTING SCORES TO FAILING EXAMINEES: SURVEY OF PROCEDURES IN USE AT COPA 18

Selection of Sample 18
Classification of Score Reporting Systems 19
Reporting of Total Test Results 24

VI. INTERVIEWS WITH MANAGERS ABOUT CANDIDATE SCORE INFORMATION 26

Summary 26
Interview Method and Procedure 26
Findings 27
How CSI Should be Improved 28
How Clients Could Benefit from Improvements of CSI 30
Questions Most Frequently Asked by Candidates about CSI 32
Surveys of Failing Candidates 34

(continued)

VII. SUGGESTIONS FROM ETS STAFF FAMILIAR WITH PROBLEMS OF AUDITING SCORE INTERPRETATION GUIDELINE 1 35

Concept of Score Recipient 36
Concept of Useful Score Interpretive Information 36
Reporting of Total Scores 36
Reporting of Subtest Results 37
Success of Repeat Examinees 37
Differences in Subtest Difficulties 38
Other Suggestions for Improving Usefulness of CSI 40

VIII. RECOMMENDATIONS FOR THE INTERPRETATION AND APPLICATION OF SCORE INTERPRETATION GUIDELINE 1 42

Score Interpretive Information 42
Bulletins of Information 43
Score Reports 43
Score Report Inserts 43
Score Recipient and Category of Score Recipient 44
Understandable 45
Useful 46
Task 1: Determine How Far Examinees Scored Below Passing 46
Task 2: Interpret Content-Area Descriptions 47
Task 3: Identify Content-Area Weaknesses and Strengths .. 49

IX. DESCRIPTIONS OF EXEMPLARY SCORE INTERPRETIVE INFORMATION 59

Scenario 1 59
Description of Test 59
Some Test Development Steps to Facilitate Score Reporting 60
Descriptions of Content 61
Development of an Instructional Package 61
Establishment of Passing Scores 61
Reporting Total Test Results 62
Reporting Subtest Results 63
Prescriptive Reading Lists for Failing Candidates 65
Scenario 2 67
Description of Test 67
A Matter of Literacy 67
Some Test Development Steps to Facilitate Score Reporting 67
Concurrent Validity 68
Establishment of Passing Scores 68
Test-taking Aids 69
Sample Tests 70
Reporting Total Test Results 70
Reporting Subtest Results 71

(continue)

X. PROVIDING USEFUL SCORE INTERPRETIVE INFORMATION TO FAILING EXAMINEES: A LIST OF RECOMMENDATIONS FOR PROGRAM DIRECTORS, TEST DEVELOPERS AND MEASUREMENT STATISTICIANS 75

Test Planning 75
Test Specifications 75
Test Development 76
Score Interpretive Information 76
Scoring of Subtests 77

APPENDICES

APPENDIX A: LIST OF TESTING PROGRAMS AND MANAGERS INCLUDED IN SCORE REPORTING SURVEY AND AND INTERVIEW STUDY 79

APPENDIX B: MEMORANDUM SENT TO MANAGERS PRIOR TO THEIR PARTICIPATION IN THE INTERVIEW STUDY 81

APPENDIX C: INTERVIEW SCHEDULE USED DURING INTERVIEWS WITH MANAGERS 82

BIBLIOGRAPHY 85

I. OVERVIEW OF THE PROBLEM

Background

Educational Testing Service (ETS) Standards for Quality and Fairness are "...designed to ensure that ETS products and services meet demonstrated criteria with respect to seven areas of basic importance." (ETS, 1983, p. iii) These areas are as follows:

1. Accountability
2. Confidentiality of Data
3. Product Accuracy and Timeliness
4. Research and Development
5. Tests and Measurement
6. Test Use
7. Public Information

Approximately every three years, ETS programs are audited for compliance to Standards, under the supervision of the Office of Corporate Quality Assurance. Auditors include ETS staff members and external personnel (e.g., college professors) assigned according to their fields of expertise. For example, a psychometrican might audit programs with respect to the reliability guidelines of the Tests and Measurement area, while a manager of testing programs might audit programs with respect to the Accountability guidelines.

Score Interpretation Procedural Guideline 1

The present report focuses on Score Interpretation Procedural Guideline 1 of the Tests and Measurement area. Guideline 1 was revised in 1983 and presently reads:

> "1. Provide score interpretation information for all score recipients in terms that are understandable and useful to each category of score recipient." (ETS, 1983, p. 18).

One noteworthy difference between the present and earlier versions of Guideline 1 is that the present version specifies that score interpretation information provided to score recipients should be "useful."

A Matter of Interpretation

During the 1982-1983 audit process, the Office of Corporate Quality Assurance noted that auditors found the new version of Guideline 1 somewhat ambiguous, especially the word "useful." Since this statement was the primary reference to be used by auditors in making their judgments, it soon became evident that, at least in the short term, evaluation of programs on Guideline 1 was not proceeding as straightforwardly as desired. Ironically, interpretation problems seemed to be most evident for programs where score recipient categories were most clearly defined: programs that administered PASS/FAIL tests. Although these programs seemed to be

particularly applicable to Guideline 1, given the availability of score recipient categories--PASS and FAIL, interpretation questions such as the following were raised by auditors and program directors:

- What is meant by the term "useful?"
- What is meant by the term "score recipient?" Would score recipients include clients as well as test-takers?
- For some programs, categories of score recipients are PASS and FAIL. What information would be useful to each group?
- What criteria should be used by auditors to determine the extent to which score interpretation information reported to failing score recipients would be "useful?"

The Office of Corporate Quality Assurance recognized that the present version of Score Interpretation Procedural Guideline 1 should be further clarified. Without sufficient clarification, not only would the auditors and program direction staff continue to experience difficulties in arriving at judgments about program compliance with this guideline, but the value of the judgments could become questionable. How valuable could a judgment be if the judgment task was not clearly understood by the judges? Moreover, since an important purpose of the Standards is to assist program direction in improving the quality and performance of their programs, how useful would such judgments be in delineating actions to be taken in effecting these improvements?

Purpose and Scope of Study

Since the scope of this problem is potentially too wide to be accommodated by available fiscal resources, the present study only addresses Score Interpretation Guideline 1 as it would apply to test-takers who fail PASS/FAIL tests. This special case is of immediate priority since failing test-takers comprise a score-recipient category with a crucial need for score interpretive information. Failure to address this need would probably result in greater consequences for failing test-takers than for other categories of score recipients (e.g., passing test-takers). Although the need to provide useful score interpretive information to other categories of score recipients is also important, both priorities and resources warrant that failing test-takers be the primary concern of the present project.

Of the PASS/FAIL tests administered by ETS, more than 90% are administered by the Center for Occupational and Professional Assessment (COPA) division of ETS. Most COPA tests are used to classify licensure and certification candidates into score categories of PASS and FAIL. Although the study focuses largely on COPA programs, it is the author's hope that the report will be applicable to all ETS programs that administer PASS/FAIL tests.

The intended readers of the report are persons participating in the ETS audit process: internal and external auditors, program direction staff and Office of Corporate Quality Assurance staff members. The document is intended to assist them with the preparation of their respective audit reports, with respect to Score

Interpretation Guideline 1. It is hoped that the present project will serve as a useful reference for Guideline 1, clarifying the interpretation of this guideline and improving communication among those involved in the audit process. The project goals are:

1. To review research literature on providing interpretive information to failing test-takers.

2. To review current program efforts to provide interpretive information to failing test-takers.

3. To develop criteria for determining the "usefulness" of interpretive information provided to failing test-takers.

4. To recommend new or revised methods for providing "useful" interpretive information to failing test-takers.

As the wording of the goals above suggests, this study first examines the efforts of ETS programs to comply with Score Interpretation Guideline 1 then makes evaluations and recommendations for improvement.

Data sources consist largely of interviews with ETS staff members--especially auditors with experience in evaluating the Score Interpretation Guidelines, recent ETS audit reports and related correspondence, written descriptions of methods currently in use at ETS and relevant scientific literature .

II. DISTINCTION BETWEEN CRITERION-REFERENCED AND DIAGNOSTIC TESTS

An important distinction should be made between criterion-referenced tests and diagnostic tests. Simpson and Arnold (1983) in their review of the literature indicated that these two terms have often been used interchangeably by persons in the educational field. Yet, the two types of tests are very different, both in their uses and in their designs (Nitko, 1980 and Simpson and Arnold, 1983).

Criterion-referenced tests (CRT's) are used to measure an examinee's status in reference to a well-defined behavior domain (Popham, 1978). Status usually refers to score categories of PASS or FAIL; examinees earning total test scores at or above the passing score would pass while examinees scoring below the passing score would fail. Although tests of this type may be based on behavior domains ranging from undefined to well-defined and ordered, all may be classified as CRT's, provided that they conform to the following broad definition:

> "A criterion-referenced test is one that is deliberately constructed to yield measurements that are directly interpretable in terms of specified performance standards." (Glaser and Nitko, 1971, p.563)

Tests of this type administered by ETS are based on behavior domains that may vary from ill-defined to well-defined. In accordance with Nitko's classification scheme (Nitko, 1980), these ETS tests will simply be referred to as CRT's (as opposed to mastery tests,

domain-referenced tests, objective-referenced tests or competency tests, etc.).

CRT's are often used for diagnostic purposes. To provide failing examinees with more specific information (than that provided by total test scores), subtest or objective scores measuring specific content-area performance could be calculated, usually by tallying the scores for items assigned to each content area. These values would usually be reported as number right and/or percent of subtest or objective items correct (Hambleton, Gorth, and O'Reilly, 1973; and Hambleton and Eignor, 1979). If passing scores have been defined for the subtests, then examinees can be told which content areas they should emphasize in preparing for retesting. It should be noted that CRT's are designed primarily to assess whether examinees have attained an acceptable status (i.e., PASS), typically in reference to the trait measured by the total test. To fulfill this purpose, CRT's are constructed to maximize the reliability (i.e., consistency of pass-fail classifications) and interpretability of the total test scores. Although CRT subtest scores may provide examinees with a general indication of subject matter strengths and weaknesses, they tend to be ineffective at revealing causes underlying failure (e.g., deficiencies in instruction). Subtest scores might indicate which broad skill areas should be emphasized in preparing for retesting, but would not indicate specific skill failures or suggest learning strategies.

Diagnostic tests, on the other hand, are especially constructed to test for the presence or absence of specific factors known to impede learning, e.g., failure to recall facts (Simpson and Arnold, 1983). These factors would have been identified previously by expert instructors through investigations of skill failures, subject content and teaching practices. As Nitko (1980) has suggested, diagnostic tests, by implication, are closely linked to instruction, since they provide means of tailoring instruction to skill weaknesses identified by the tests.

Diagnostic tests can be constructed to elicit erroneous responses that could be traced to improper learning or instruction. For example, in a multiplication test, the test constructor would deliberately include items that would require the operation of "carrying." Each item would measure a skill failure to be scored as 1 (skill failure is indicated by response) or 0 (skill failure is not indicated by response). The items measuring a class of skill failure (e.g., carrying incorrectly) would be scored as a cluster, by tallying the item scores (Gropper, 1975). Examinees scoring low on the cluster would be provided remedial instruction in carrying operations. Thus, not only could a cause of deficiencies be identified, but instructional strategies could also be suggested.

While CRT's emphasize assignment of examinees to score categories (e.g, PASS/FAIL), diagnostic tests emphasize the identification of specific skill failures for the purpose of prescribing remedial action. It should not be surprising that neither type of test would be as effective at serving the secondary purpose. Some of the confusion between these two types of tests is

probably due to the fact that to most nondiagnosticians they appear to be the same in form and content. Indeed, diagnostic tests and CRT's based on the same content may look the same to the untrained eye. Further confusion may result from the fact that experts such as Gropper (1975) have advocated replacing some items on a CRT with diagnostic items. The reasoning is that the diagnostic usefulness of CRT's would be improved, though possibly at the cost of the reliability of both the diagnostic and CRT scores. A more acceptable practice is to first administer a test (usually a CRT) that can identify learning weaknesses in a broad way, then administer a diagnostic test to obtain more specific information about skill failures (Payne, 1974; Simpson and Arnold, 1983; and Spencer and Smith, 1969). The use of CRT's for diagnostic purposes is limited largely by the construction and purpose of CRT's. By recognizing the limited diagnostic usefulness of CRT's, one may better understand how much the diagnostic usefulness of CRT's can be improved without jeopardizing their main function.

III. REPORTING OF CRITERION-REFERENCED TEST RESULTS

In their review of the literature, Mills and Hambleton (1980) found a scarcity of research literature on the reporting of criterion-referenced test results. The authors considered this finding unfortunate since:

> "...the purpose of a testing program is, after all, to provide usable test score information (to appropriate audiences) in a convenient format." (Mills and Hambleton, 1980, p. 1)

Test score information that is ambiguous or inappropriate would be of limited value to potential users.

In addition to reporting scores as total scores (including pass/fail status), scores are also reported in variously defined subcategories, depending on the assessment strategy used. Major purposes of such subscores are to provide examinees, instructors and administrators with information to aid them in interpreting total test status and to offer some diagnostic guidance.

Millman (1978) identified three general strategies for constructing CRT's:

1. Item-centered strategy
2. Subtest-centered strategy
3. Objective-centered strategy

In the item-centered strategy, a test may cover a large number of objectives, with each objective being measured by as few as one or two items. This strategy makes it possible to evaluate examinees on a wide variety of objectives, perhaps hundreds, although the reliability of each evaluation may be questionable. This strategy is used by the National Assessment of Educational Progress.

The subtest-centered strategy involves constructing a test to measure 5-10 broad skill areas. Each skill area is usually measured by a sufficient number of items to be interpretated as a single, reliable subtest. However, since each subtest may measure several objectives, the content descriptions of the subtests may be difficult to clarify for examinees and instructors. Of the three strategies considered, this is the strategy used most frequently at ETS in constructing CRT's.

The objective-centered strategy is the preferred assessment strategy for diagnostic uses of CRT's (Ahmann, 1978). With this strategy, a CRT includes no more than 20 objectives, each measured by a number of items sufficient for an acceptable level of reliability--a minimum of about 15 items (Haladyna and Roid, 1980). This strategy is preferred also because it is conducive to the use of well-defined objectives, which benefit test score interpretation and test construction. A disadvantage of this strategy is that the number of objectives included within a single test is limited.

In a California Department of Consumer Affairs memorandum (C. Foutz, personal communication, November 30, 1984) issued to state licensing boards, the Chief Deputy Director recommended that

diagnostic services be provided to examinees who fail licensing tests. The document stressed that, at a minimum, failing examinees should receive pass-fail decisions for subtests as well as for the total test; licensing boards were also advised to report, at their discretion, total test and subtest scores (if the subtest scores are reliable), and appropriate score interpretive information. Among the points emphasized by the Chief Deputy Director is the following:

> "Helping failed examinees identify their deficiencies in knowledge or skill is the single best way to preclude formal examination complaints, litigation, and demands for full disclosure of the test." (C. Foutz, personal communication, November 30, 1984, p.3).

IV. WHAT DIAGNOSTIC INFORMATION CAN BE PROVIDED BY CRITERION-REFERENCED TESTS?

Although CRT's may not provide diagnostic information to an acceptable degree (Simpson and Arnold, 1983 and Spencer and Smith, 1969), CRT's can provide some information that may be considered useful to score recipients in interpreting test scores. Most of the relevant literature focuses on CRT score interpretations within school settings. Mills and Hambleton (1980) identified information that should be available in reports of CRT scores for four audiences: (1) teachers, (2) parents and students, (3) school building administrators and (4) district administrators. For each audience, primarily at the elementary and secondary school levels, the authors listed the information needs and the use of the information. For example, for parents and students, one information need might be identification of specific objectives for which performance is low. The appropriate use of score information might be to determine specific deficiencies. Some other ways to meet the informational needs of parents and students are: inclusion of sample items from nonmastered objectives, narrative diagnostic and interpretive reports to supplement numerical summaries and performance relative to other students in the same class. It is unfortunate that the authors did not detail how much or what information would be provided to the audiences in these reports. For example, what information should go into a diagnostic report and how should it be derived? Nevertheless, these guidelines represent the most ambitious attempt at establishing guidelines for CRT

reporting that this author has found in the scientific literature.

One CRT reporting system cited by Mills and Hambleton as providing useful information, especially at the classroom level, has been described by Furlong and Miller (1978). This computer-based system, DIAGNOSE, provides summaries separately for students and teachers at the college level. The summaries are nontechnical and reasonably simple to interpret. Students at selected score categories (e.g., students earning failing scores) receive a list of the questions they answered incorrectly, showing their incorrect responses and the correct answers. In addition, students may receive (at the teachers' option) a description of the objective to which each incorrectly answered item was assigned and a list of instructional materials for each nonmastered objective.

The report to teachers may include (at the teachers' option) summaries of total group performance (apparently raw total test scores) by course objectives, by critical dimensions of curriculum, or by taxonomic level of objectives (e.g., comprehension). Also reported is the percentage of students passing all items within each objective and/or taxonomic level of objective. For test and item evaluation, numbers of students selecting each item-choice and total test statistics (means, standard deviations, etc.) are provided.

It should be noted that at the student level especially, Furlong and Miller's reporting system is straightforward. Students are provided a copy of the items they answered incorrectly, along with the correct answers to these items. They then know exactly how many additional items they would have needed to answer correctly to

pass the test. Moreover, they are told which course objectives they should study and which instructional materials to read. This system does appear to provide students, especially failing students, with useful information to help them improve their scores. On the surface this appears to be a good example of how CRT's can be used diagnostically. McArthur and Choppin (1984), however, reported that teachers encountered much difficulty interpreting and using the results correctly during the brief period when DIAGNOSE was operational at Southern Illinois University. This author learned through inquiries that the teachers did not appear to understand the diagnoses well enough to explain them to their students; in addition, they indicated that they had to spend an inordinate amount of time preparing for and reporting individual diagnoses to students.

Another cautionary note is offered for readers who might consider applying Furlong and Miller's reporting system to other settings. While the disclosure of answer keys and test items may not present a serious problem in classroom settings, where tests are constructed quickly and cheaply, such disclosures could be prohibitive for large testing corporations such as ETS. If answer keys are disclosed, new items have to be written to replace the compromised (disclosed) items, in order to maintain acceptable levels of reliability and validity for the test scores. Since many of the CRT's administered by ETS are based on behavioral domains that are limited in scope, the possible number of items that can be written is finite. How many state laws should a candidate for licensure as a real estate salesperson be required to know? How

many items can be written to cover such a narrow, though important, domain without inducing excess trivia or redundancy? If Furlong and Miller's system is to be applied to CRT's administered by ETS, significant adjustments would have to be made.

Percent-Correct Scores

More typically, CRT subtest results are reported as percent-correct scores (Hambleton and Eignor, 1979 and Wilson and Hiscox, 1984). Although Hambleton and Eignor seemed to prefer a Bayesian-derived mastery index as an alternative (to percent-correct scores), they considered percent-correct scores to be a reasonable way to present individual data.

Wilson and Hiscox (1984) recommended that school district personnel use standardized test results to derive average percent-correct subtest or learning-objective scores that closely match local school curricula. To calculate what the authors termed "average adjusted scores," recently tested school districts must first obtain, for all items on the standardized test, detailed content descriptions and the proportion of school-district students who correctly answered each item (p-value). The items are regrouped into subtests or objectives that closely match skills actually being taught in the schools. The p-values assigned to these newly defined subtests or objectives are summed, divided by the number of subtest or objective items and finally multiplied by 100. The resulting "average adjusted scores" are merely mean percent-correct scores. Wilson and Hiscox considered percent-correct subtest and learning

objective scores to be useful for identifying skill strengths and weaknesses, both at appropriate aggregate levels (e.g., school level) and at the individual examinee level.

Rather than provide examinees with answer keys and their item responses, percent-correct scores could indicate how close failing examinees came to the passing scores both in terms of the subtest or objective and the total test. If subtests or objectives are well defined and accompanied by sample items for each subtest or objective, failing candidates should be able to apportion their study efforts according to skill weaknesses. The disadvantages and advantages of reporting percent-correct scores will be discussed in more detail in the next section, within the context of ETS administered CRT's.

V. REPORTING SCORES TO FAILING EXAMINEES: SURVEY OF PROCEDURES IN USE AT COPA

Selection of Sample

In January 1984, a sample was drawn from all Center for Professional and Occupational Assessment (COPA) licensure and certification testing programs. Included in the sample were programs that :

1. Administered criterion-referenced tests,
2. Were fully operational (as opposed to new or terminating programs) and
3. Had their candidate score reports issued by COPA (rather than another organization).

Of the 50 programs in operation at COPA at the time of the survey, 43 were testing programs; 23 programs fulfilled the sampling requirements and were selected for the study. The 23 testing programs surveyed are listed in Appendix A. Readers should be aware that the findings of this study will not be reported by individual program.

Classification of Score Reporting Systems

After reading score-interpretive documents (candidate bulletins, score reports, etc.) written for each program in the study, in addition to interviewing program direction staff, this author determined that the types of diagnostic information provided by COPA testing programs could be classified into three broad categories:

1. Diagnostic (subtest) scores reported as either number or percent of subtest items right
2. Diagnostic results reported as level of performance (e.g, "Fair," "Good," etc.)
3. No diagnostic information reported

Number- or percent-right scores were, for every applicable program, raw scores reported for each of several diagnostic categories. The rationale for this is that candidates will be able to identify the content areas in which they were strongest or weakest. The information is usually targeted to failing candidates, who may then use the results to determine which content areas they should emphasize in studying for retesting and which areas they can de-emphasize. Although the diagnostic categories are typically described in candidate bulletins, no cut-off scores are provided for each diagnostic score. Candidates therefore have to determine for themselves how low their diagnostic scores must be to warrant concentrated study. In addition, there is the problem of diagnostic subtests not being of equal difficulty. The problem then arises of

applying diagnostic scores earned on one test form to corresponding diagnostic scores earned on another test form assigned to the same candidate for retesting. However, number- and percent-right scores offer the advantage of simplicity, in that diagnostic scores are in a form with which the candidates are very familiar--number or percentage of test items answered correctly. And, as noted in the literature review, this method has been considered very useful for this purpose.

"Performance Levels" were verbal statements that invariably ranged from the lowest level of performance (e.g., "considerably below the minimum competence") to the highest level of performance (e.g., "considerably above the minimum competence level"). This method offers a clear advantage over the number/percent right score method, in that candidates are not forced to determine for themselves how well they performed on each diagnostic subtest. The levels are defined statistically by subtest score ranges as percent-correct, as illustrated by the following example taken from an actual score report:

Diagnostic Category	% of Questions Correctly Answered
Excellent	96% and Above
Good	86% to 95%
Fair	76% to 85%
Marginal	70% to 75%
Needs Improvement	69% and Below

This example is typical, in that a passing score range (70% to 75% and above) is defined for all subtests and test forms. The passing-score ranges usually have been identified through past experience: "Passing scores (as percent-correct) reported for the total test usually fall between XX% and XX%." If subscores are perfectly positively correlated with total scores, and if passing scores actually fall within the marginal range, candidates should be able to use their diagnostic test results to help them to understand their total test scores. For example, a candidate who earns a low passing score on the total test might receive subtest results of "Marginal" (Subtest 1), "Needs Improvement" (Subtest 2), "Marginal" (Subtest 3), and "Marginal" (Subtest 4). The candidate's overall performance on the subtests is marginal, with some weakness in the content area measured by subtest 2. Such subtest results would seem to make sense, because of their consistency with the total test score.

Some candidates may receive diagnostic results, however, that could be considered perplexing. This is primarily due to the fact that reported passing scores (as percent-correct) have occasionally been observed to drift outside pre-fixed "marginal" ranges. For example, a candidate who fails the total test by one percentage point at a passing score of 77%, might receive a result of "Fair" for all or most subtests. It is doubtful that this outcome would be useful to the failing candidate who hopes to identify content areas to emphasize in studying for retesting. (Note: Since reported passing scores are usually derived through linear score equating methods, the raw score equivalents tend to vary from one test form

to the next, depending upon the degree of score adjustments required to maintain a constant reporting score scale.)

The author has proposed a method of improving the consistency between total and subtest results. In this method, reported (equated) total-test passing scores would be converted to a percent-correct score. This score, in turn, would be applied as a standard for each subtest to obtain subtest cut scores expressed as percent-correct. The resulting percent-correct scores would be assigned to performance levels defined in units of the standard error of measurement from the calculated subtest passing scores. Lord (1984) discusses four methods for approximating the standard error of measurement at specified passing scores. The subtest results would be reported with verbal descriptions, e.g.,"Competency Not Demonstrated, Although Marginal: Some Additional Study is Recommended." Since the calculated subtest passing scores would be mathematically related to the total test scores, the diagnostic results reported to candidates would be consistent with their overall pass/fail status. This condition would be maintained even when reported passing scores (as percent-correct) vary widely. This method should only be used if the following two conditions exist: correlations between subtest and total scores are high and positive, and subtests are of comparable difficulty.

Use of these diagnostic reporting modes by the 23 programs is distributed as follows:

	Programs Assigned to Category	
	Number	Percentage
Category 1: Number or Percent Right Scores	11	47.8%
Category 2: Performance Levels	10	43.5%
Category 3: No Information	2	8.7%
Totals:	23	100.0%

These results indicate that, of the programs reporting diagnostic information, about half reported diagnostic results as number- or percent-correct scores and about half reported these results as performance levels. More than 90% of the COPA testing programs surveyed made at least some effort to provide diagnostic information to failing candidates. Whenever diagnostic information was reported, failing candidates received it; a few programs also provided diagnostic information to passing candidates. Only two of the 23 programs surveyed did not report any diagnostic information to candidates.

Reporting of Total Test Results

According to program direction staff interviewed, one of the questions frequently asked by failing candidates is: "How close did I come to passing the test?" Usually descriptions of the score scale are provided to candidates in bulletins of information, which include explanations of how scores are derived, the defined range of the scores and the passing score. In cases where total scores are reported, typically only to failing examinees, candidates obtain some information about how close they came to earning a passing score. In other cases, however, although explanations of score scales are provided, candidates are merely informed of whether they passed or failed; a statement to that effect, sometimes with as little as "P" for passed or "F" for failed, appear on either a score report or a letter. Although failing candidates almost always receive results of their performance on diagnostic subtests, their question ("How close did I come to passing the test?") is rarely derivable from the information provided. The test programs in the sample reported total score results as follows:

Total Test Result Reported to Failing Candidates	Testing Programs in Sample (N=23) Number	Percentage
Test Score:	17	73.9%
Pass/Fail Status:	6	26.1%
	---	------
Totals:	23	100.0%

This indicates that the practice of reporting test scores to failing candidates was prevalent among the programs sampled. Six (26%) programs, however, reported total test results only as pass/fail status.

VI. INTERVIEWS WITH MANAGERS ABOUT CANDIDATE SCORE INFORMATION

Summary

The managers of the 23 testing programs in the analysis commented on the candidate score information (CSI) currently provided to failing candidates and identified improvements they would like to see made. The managers also identified many improvements they would like to see made to other CSI currently being used for their programs. Foremost among their suggestions was the need to tighten the definitions of test content areas and to develop content-area scores consistent with candidates' total-score results. In addition, the managers identified questions candidates ask most frequently about the interpretation of their score results. These questions can provide useful information for improving current and designing new CSI. When asked about the possibility of surveying failing candidates about CSI, most of the managers felt such an effort could lead to improvements of CSI, and that most of their clients would consent to such a survey.

Interview Method and Procedure

COPA program/operations directors were viewed as important sources of information about failing candidates' problems and questions regarding candidate score information (CSI), the information sent to candidates for the purpose of reporting or explaining test results. Interviews were conducted with program and

operations directors of all 23 COPA testing programs in the analysis sample. At least one program/operations director from each of the 23 programs in the sample was interviewed by the author, resulting in a total of 6 interviewees; all interviewees were responsible for two or more programs. A list of the participants and their programs appears in Appendix A. Prior to the interviews, the managers each received a memorandum (Appendix B) that provided some background information about the project, the general purpose of the interview, a list of their testing programs that were to be the focus of the interview (i.e., those in the analysis sample) and a list of the actual interview questions to be asked. The interviews were conducted between May 8 and 29 of 1984, and were about 20 minutes in duration. A copy of the interview schedule appears in Appendix C.

Findings

Described below are noteworthy concerns, opinions and recommendations of the 6 managers, with respect to CSI, listed in order of importance in the author's judgment. The "N" that appears in parenthesis next to each response is the number of managers (out of 6) who responded in that way.

How CSI Should be Improved

- "Tighten the definitions of test content areas." (N=2)

  Shortcomings of existing content areas seem to surface when failing candidates request more specific information from clients and ETS about what topics they should study to prepare for retesting. Unless the content measured by a particular subtest score is interpretable to failing candidates, no statistical procedure for deriving subtest scores will be useful to the candidates.

- "Develop content-area scores that are consistent with candidates' total test results." (N=4)

  Problems with inconsistency between subscores and total scores are most prevalent among candidates who score just below the passing score. Interpretation difficulties arise when these borderline failures receive content-area results that seem to contradict the total test result. For example, a candidate may sometimes receive the content result "Fair" for all content areas and still receive a failing result on the total test.

- "Simplify explanations of scaled scores that appear in candidate bulletins of information and score-report inserts." (N=4)

Some candidates interpret scaled scoring to be the same as grading candidates on a "curve," or percent-correct scoring. For more than 80% of the testing programs sampled, these interpretive problems can be alleviated by making the appropriate clarifying revisions to candidate bulletins and score-report inserts.

In 4 (17%) of the testing programs sampled, managers have observed that candidates frequently experience difficulties reading even the most simply worded sections of candidate bulletins. The programs of concern are certification testing programs in vocational-technical areas such as cosmetology or construction code inspection. While test scores derived through complicated means such as equating certainly offer advantages of precision (e.g., minimum variation from the original passing score regardless of the difficulty of new test forms), they may confuse academically unsophisticated candidates to the point that problems of interpretability of test scores may outweigh the benefits of precision or the benefits of providing this information in CSI. Concerns of this nature suggest that methods of deriving test scores should be determined with consideration to the level of understanding of candidates--especially failing candidates. Managers and psychometricians should ask themselves: "Can the types of test scores under consideration be explained clearly to failing candidates?"

This is not an argument for imprecision; the first duty of a psychometrician is to maximize precision of measurement. Rather, what is offered are thoughts to consider when the choices among scaling/equating methods differ little in degree of precision.

How Clients Could Benefit from Improvements of CSI

- "Improved CSI may reduce the number of questions clients would receive from candidates about score reports." (N=4)

CSI that clearly explains scoring concepts (e.g., equating) and is perceived by candidates as useful in preparing for retesting would reduce the number of questions that candidates ask of clients. The managers indicated that, for various reasons (apprehension about being sued, uncertainty about how to respond to candidates, interference with clients own work schedules, etc.), clients wish to reduce the number of candidate inquiries to a minimum.

- "Improved CSI may reduce number of legal challenges." (N=1)

Occasionally, failing candidates have threatened to take legal action against clients because they (failing candidates) have misinterpreted candidate score information. For example,

candidates may infer that their total test results should have been other than that reported, because of their subtest results. Well-written score interpretation information would clearly explain that these two scores are derived and interpreted differently.

- "Improved CSI may enable clients to appear (to candidates) to be more knowledgeable about the technical aspects of the testing program." (N=2)

  Improved CSI would be easier to understand. This would make it easier for clients to grasp the subject matter and be better prepared to respond intelligently to candidates' questions.

- "Improved CSI may enhance public relations between clients and candidates." (N=2)

  When candidates become aware that efforts are being made to improve the usefulness of score information, they will probably perceive this as an expression of clients' interest in providing candidates with a quality product.

Questions Most Frequently Asked by Candidates about CSI

- How many questions did I answer correctly?
- How many questions did I have to answer correctly to pass the test?
- Why did I need to answer more questions correctly to pass this test (form) than I needed to pass the previous test (form)?
- What is the standard error of measurement?
- Since my score was less than one standard error of measurement below the passing score, shouldn't I have passed the test?
- Which questions did I miss?
- What is equating?
- What is scaling?
- What percentage of candidates passed (or failed) this test?
- What should I study to prepare for retesting?
- What reading materials should I use to prepare for retesting?
- You mean I wasn't graded on a curve?
- What does this content description mean?

- How could I have failed the total test when I scored well on the subtests?
- Where did my actual subtest score fall within the subtest score category?
- Are the item responses weighted?
- You mean these numbers (i.e., percent-correct scores) aren't percentiles?

Nearly all candidate questions about CSI are asked by failing candidates. For all programs in the sample, usually less than 1% of failing candidates contacted ETS or client organizations; the number of candidate inquiries increases if the CSI is new or unclearly described. For the most part, candidates' questions reflect the difficulty they experience with understanding concepts of scaled and equated scores. Through their questions about CSI, candidates attempt to reduce these concepts to terms with which they feel comfortable, most notably raw test scores.

One of the managers mentioned that some candidates seemed to believe that client organizations are using scaling and equating to conceal practices detrimental to candidates, e.g., limiting the number of candidates who pass in accordance with the job market by manipulating the raw passing score. Such perceptions by candidates can weaken the trust they have in client organizations. These problems might be avoided by providing candidates with descriptions of reported scores that are simply, clearly and comprehensively

written; descriptions that implicitly reassure candidates that clients are honest and fair, and have the candidates' best interests in mind, to the extent that the integrity of the profession is not compromised.

Surveys of Failing Candidates

No surveys of failing candidates had ever been conducted for any of the 23 testing programs to assess these candidates' evaluations of the usefulness and clarity of CSI, according to the managers interviewed. The term "survey" refers to the systematic collection of facts, opinions and comments from people sampled from a target population of interest (e.g., all people registered to vote in New Jersey in November 1984), through questionnaire mailings, telephone interviews or other methods.

However, the managers of 18, or 78%, of the 23 testing programs felt their programs would benefit from this type of survey. They also felt that clients of 65% of the programs would consent to a survey of this type. While the worth of conducting survey studies of failing candidates was acknowledged by the managers, personnel and fiscal resources were insufficient for the task; in addition, the question of confidentiality of test scores had to be considered.

VII. SUGGESTIONS FROM ETS STAFF FAMILIAR WITH PROBLEMS OF AUDITING SCORE INTERPRETATION GUIDELINE 1

In November 1983, the author met separately with four ETS staff members experienced with problems encountered when testing programs were audited on Score Interpretation Guideline 1. The conversations with the ETS staffers were informal and open-ended. In general, the conversations were quite informative and insightful. The participants were:

Robert F. Boldt
Senior Research Scientist
Measurement Research and Services

Thomas F. Donlon
Program Research Scientist
College Board Statistical Analysis

Gary L. Marco
Director of Statistical Analysis
College Board Statistical Analysis

S. Donald Melville
Executive Director
Reference Services

All of the staffers indicated that Score Interpretation Guideline 1 should be clarified. In particular, they cited the terms "score recipient" and "useful" as expressions that should be clarified.

Concept of Score Recipient

As the guideline is stated, the term "score recipient" can refer to almost any person or organization. One of the staffers felt that the Score Interpretation Guidelines should be restricted to two groups of score recipients: examinees and clients. He argued that the score interpretation needs of these two groups of score recipients are different and that Score Interpretation Guidelines should be defined and listed separately for each group.

Concept of Useful Score Interpretive Information

Reporting of Total Scores

All staffers felt that testing programs that report only "pass" or "fail" to examinees do not provide sufficient information to failing examinees about how close their score was to the passing score. They recommended that total test scaled scores be reported on a continuous scale, with an identified range, and that examinees be provided with appropriate scaled passing scores.

The author concurs with this position, while emphasizing that total test scores should not be reported to passing examinees. Such scores could be misused to compare or rank examinees, e.g., to select among examinees in order to fill job vacancies. Criterion-referenced tests should not be assembled or calibrated to maximize differences between scores, as is the case with norm-referenced tests. Rather they should be designed to measure

performance accurately and validly at one score point: the passing score.

Reporting of Subtest Results

All staffers interviewed specified that some form of subtest score results should be reported to failing examinees. The typically reported percentage- or number- correct subtest scores seemed to be minimally sufficient to the staffers. However, staffers suggested ways to improve the usefulness of raw subscores, which will be discussed below.

Success of Repeat Examinees

When asked to indicate score interpretive information failing examinees would consider useful, Boldt, Donlon and Marco indicated that failing examinees should be provided with some information about the success of repeat examinees at passing retests. A general approach offered by these three staffers involves first accumulating data on repeat test-takers for several test administrations (until a sufficient sample of examinees has been acquired). According to Boldt, the cumulative scaled test scores could then be used to build expectancy tables indicating the probability of passing a retest given specified score ranges on the initial test. Failing examinees would use their recently reported scores to look up the corresponding probability of earning a passing score at the next testing. The interpretation (by candidates) of these tables may be improved by simply reporting a record of past repeat candidates' performance (e.g., passing rate on the first retesting). Users of

this approach must identify and control (experimentally or statistically) factors which operate on repeat candidates' test performance. Even in laboratory experiments on human subjects, intervening variables can be too numerous and nebulous to be tallied, let alone measured and controlled. In realistic settings, the task of obtaining interpretable measures of repeat candidate test-retest performance often borders on the impossible. This is certainly true for COPA licensure and certification testing programs, where almost all examinees are adults who are free to prepare--or not prepare--for reexaminations in any way they might choose.

Differences in Subtest Difficulties

Donlon and Marco addressed an equating problem that has been cited by auditors of the Score Interpretation Guidelines over a period of several years. The problem is that subtest results that are reported have rarely been adjusted for differences in difficulty, that is, have rarely been converted to equivalent scales. As a consequence, a percent-correct score of 50% on one subtest may correspond to a score of 60% on an easier subtest, or to a score of 40% on a harder subtest of the same total test. Examinees who receive such unadjusted subtest scores most likely interpret them as though they were all on comparable scales, thinking that a percentage score on one subtest represents the same level of performance as an identical percentage score on any other subtest. This is misleading, especially if the subtests differ widely in difficulty.

One solution offered by Donlon was to advise examinees about the problems in interpreting scores for subtests that differ in difficulty, though only when differences can be assumed to be slight. No statistical adjustments for difficulty would be applied to the subtest scores. The implied benefit would be that examinees would be cautioned against making oversimplified interpretations and accepting the subtest results as exact information. This approach would be feasible when the mean differences in subtest scores are known to be small and stable from one test form to another. If differences are not trivial, it would not make sense merely to inform examinees that subtests differ in difficulty; examinees should also be informed about the directions and sizes of differences.

One suggestion by Marco was to scale the subtest scores to the scale of the total test scores, usually through mean-standard deviation equating. This approach would adjust for differences in subtest difficulty by placing the subtest scores on a common scale, provided that the total-test-score distribution is linearly related to all of the subtest score distributions. (If the analysis sample is large; i.e., 1,000 or more examinees, equi-percentile equating would be preferable to mean-standard deviation equating since relationships between total test and subtest score distributions can sometimes be nonlinear.) Should this approach be used, it would be necessary to define interpretable cut scores or score ranges that indicate levels of performance. For example, a scaled subtest score of 75 or higher may indicate that examinees have demonstrated that they are competent in the content area measured by the subtest

scores. Without such guidelines, the scaled subtest scores would be even less meaningful to examinees than the raw number-right scores.

Other Suggestions for Improving Usefulness of CSI

Donlon suggested that a supplement accompany Score Interpretation Guideline 1. The supplement would serve to describe the guideline more fully, and provide hypothetical examples of effective and ineffective CSI. Each example would be analyzed in detail for the purpose of providing program staff and auditors with criteria by which to evaluate operational CSI. This suggestion is certainly workable.

Melville felt that all examinees should be provided with an annotated bibliography of reading materials that were used to write test items. While this is a fine suggestion, it may be difficult to implement in licensure and certification testing programs, since content experts tend to rely more on their experience than on text materials as a basis for writing and reviewing items. In addition, in regard to licensure programs, state laws sometimes change too rapidly for even state commissioners to keep pace.

A final suggestion for improving CSI offered by Marco was to provide examinees with practice questions that all minimally competent examinees would be expected to answer correctly. The items would be assembled into a mini-test, representative of the total-test content. Each examinee would score his or her own practice test, with the expectancy of correctly answering 100% of the questions. Thus, the cut score for each of the subtests would

be 100% as well. This approach would require that item data be available beforehand, either item ratings from standard-setting studies (e.g., Nedelsky method) or examinee score data (e.g., test scores of known masters).

This is a workable suggestion that can be improved further by providing examinees with item difficulty (i.e., P+) values based on borderline examinees' responses to the same items as they appeared on (full-sized) operational test forms. This information should enable examinees to compare their performance on the mini-test items with the performance of the examinee group for which the test was targeted. Before this suggestion can be implemented, the effects of the two test contexts (mini- vs. full-test) on item difficulty would require investigation.

VIII. RECOMMENDATIONS FOR THE INTERPRETATION AND APPLICATION OF SCORE INTERPRETATION GUIDELINE 1

The major purpose of this study is to provide further clarification of Score Interpretation Procedural Guideline 1. The focus of this study is on examinees who receive failing score reports. Guideline 1 is:

> "Provide score interpretation information for all score recipients in terms that are understandable and useful to each category of score recipient." (ETS, 1983, p. 18)

If this guideline is to continue to be worded as it currently is, a supplement should be developed for use by auditors, program directors and Office of Corporate Quality Assurance staff. The following draft supplement is offered as a starting point, with some discussion about each of the terms and phases in the guideline that appear to warrant further clarification, as underlined above.

1. Score Interpretive Information

Score interpretive information, or candidate score information, refers to all documents examinees receive regarding the interpretation of their score reports. Typically, this would

consist of:

- Bulletins of Information - Booklets designed to describe, in general terms, procedural and background information examinees are required to know or may wish to know about the testing program--this includes information about the client organization, test center locations, eligibility criteria, how to apply for testing, test content and explanations of test results. This document serves as the examinees' major information source about the testing program. Bulletins of information are usually mailed to examinees several weeks in advance of the test administration.

- Score Reports - Printed statements (sometimes in the form of a business letter) sent to examinees which indicate total-test performance (usually a scaled score or pass/fail indicator) and, generally subtest results (usually a percentage or number of items correct score).

- Score Report Inserts - Information sent to examinees with their score reports, usually as an aid to examinees in interpreting their score results--score report inserts are often used to alert examinees to changes and additions to

bulletins of information or to reiterate score interpretive information (e.g., to caution that scaled scores should not be interpreted as percent-right scores).

2. Score Recipient and Category of Score Recipient

Score recipients include persons (e.g., failing examinees) and institutions (e.g., national certification boards) who are the intended recipients of score reports; i.e., those for whom the score reports are designed. In most instances, the score recipients are examinees. Each score report is generally designed for one recipient; exceptions would probably be inappropriate since the needs of each type of score recipient tend to be different. For example, a failing examinee may be interested in knowing how far his/her score fell below the passing score, or what skills should be emphasized when preparing for retesting.

Unlike examinees, institutions such as national or regional certification or licensure boards would probably require aggregate score information, most notably, how many examinees failed the test. If the failure rate is high, an institution may become concerned that examinees will complain about the accuracy and validity of the test, and that some examinees may pursue legal action against the institution. On the other hand, if an inordinately large proportion of examinees pass, the institution may worry that examinees and professional peers may criticize the certification or licensure

process as being irresponsible and unworthy of the profession. Score information that institutions require is typically provided in statistical summary reports and examinee score rosters.

3. Understandable

Within the present context, the term "understandable" describes score interpretive information that score recipients can read and comprehend well enough to apply to their own test results. Examinees who understand score interpretive information should be able to interpret their score reports and take the appropriate actions, according to instructions. For example, after reading the score interpretive information, failing examinees may be expected to a) know that they failed, b) know how close they came to passing (in scaled score units) and c) identify content areas in which they should concentrate their studies.

This concept should not be confused with usefulness; score interpretive information may be quite useless and still be understandable. For example, a clerical error may have resulted in the content-area definitions being interchanged; i.e., the definition for content area 1 being switched with that of content area 2. Although the content definitions may be both comprehended and applied by the examinees, the definitions would be useless for identifying examinees' actual skill weaknesses. Understandability is a necessary but not a sufficient condition for usefulness. The relationship between understandability and usefulness is somewhat analogous to the relationship between reliability and validity.

4. Useful

From score interpretive information provided, failing examinees should at least be able to:

1. Determine how far their scores fell below the passing score,

2. Interpret the content-area descriptions clearly and

3. Identify content areas in which they should concentrate to prepare for retesting.

If failing examinees cannot successfully complete these three tasks using the score interpretive information provided, the score interpretive information is not useful.

Task 1 - Determine How Far Examinees Scored Below Passing

Task 1 implies that failing examinees should receive total score results as continuous scores or as polychotomous score categories (e.g., high, marginal, low). Continuous scores, usually scaled scores, would be preferable. Total-test results reported as "passed" or "failed" would not be sufficient, since failing examinees would not be able to determine the extent of their failure. It is important that descriptions of the scale (lowest and highest reported scaled score, passing scaled score, etc.) be available to examinees. An examinee who is informed that he/she has

failed by a substantial margin may be more motivated to prepare for retesting than an examinee who has failed by a smaller margin. An examinee who becomes hopelessly discouraged on receiving the mere statement "failed" might gain confidence on being informed that he/she had failed by a single scaled-score point on a 200-point scale. Feedback about the extent of failure is useful information, and should be provided to all failing examinees automatically.

It generally would not be advisable to report total-score results to passing examinees, since criterion-referenced tests are usually developed to assign examinees to pass/fail categories rather than to make comparisons among examinees. In theory, criterion-referenced tests are assembled to maximize precision at the passing scores; while the full score range is not ignored during the assembly, calibration, equating and scaling of criterion-referenced tests, it receives less emphasis.

Task 2 - Interpret Content-Area Descriptions

Task 2 requires that examinees receive detailed descriptions of content areas measured by the subtests, usually in bulletins of information. Examinees should be able to link these descriptions to study materials available to them.

This can be a problem since institutions (colleges, state agencies, etc.) who develop and teach the training courses that prepare examinees to take licensure and certification tests administered by Educational Testing Service usually operate independently from the persons who develop the tests. Test security

appears to be one of the primary reasons for keeping these two processes separated. Although the subject matter taught in the training courses and that measured by the tests tend to be somewhat similiar, the absence of a marriage of training and assessment makes it difficult to describe test content in very specific terms (e.g., behavioral objectives). Of course, auditors should be alert to detect test content expressed so broadly as to be uninterpretable by failing examinees.

In addition to content-area descriptions, a short sample test (15-20 items) and scoring key should be provided to examinees, usually in bulletins of information. The sample tests should be representative of the overall tests in both content and difficulty, and the items should be of reasonable quality. Passing scores could be approximated by a number of means, such as having content experts rate the items in accordance with an acceptable standard-setting method (e.g., Nedelsky method), or using Item Response Theory methods to build sample tests tailored to the ability level of minimally competent examinees. Thus, examinees would be in a position to test and score themselves and to determine their pass/fail status on a test very similar to one they will take under official conditions.

To further aid examinees' interpretations of scores for the test content areas, scoring specifications for the sample test should be provided, indicating assignments of items to these content areas. This should not only help examinees to refine their understanding of the content definitions, but it may also help them to discover which subtests pose the greatest difficulty for them.

Task 3 - Identify Content-Area Weaknesses and Strengths

Failing examinees attempting to complete Task 3 should be provided with subtest results that indicate how well they performed in each content area. This implies that continuous or polychotomous (score categories) numbers should be provided, along with cut scores indicating levels of strength and weakness. For example, subtest scores below 40 on a 10-80 scale may indicate examinees' performance on the content area is deficient. Regardless of the metric of the subtest scores, cut scores must be defined to guide examinees in evaluating their performance in the content areas. The practice of merely reporting subtest scores, usually raw number- or percent-correct scores, without providing examinees with interpretive guidelines (e.g., "Scores below 40 indicate that performance in this content area is deficient.") is inadequate, and should be cited by auditors.

The problem of subtests of unequal difficulty can mislead failing examinees who must complete Task 3, especially if the same cut scores are applied to all subtests. An auditor should ensure that this problem is minimal, by examining subtest means, standard deviations, intercorrelations and any other relevant statistical data, preferably for several test forms. Since the subtests probably differ in length, some simple computations could be used to convert the summary statistics to a comparable scale (e.g., percentage scale) so that subtests can be compared. Small differences among the mean difficulties of subtests can be expected because of measurement and sampling error, even if, in the true sense, the subtests are identical in difficulty. Auditors should

become concerned if mean differences are consistently marked, since failing examinees' ability to complete Task 3 successfully may be in jeopardy. If this problem occurs, program directors and measurement statisticians should be advised to have cut scores defined separately for each subtest, taking difficulty into account. This will not be feasible if the subtests consist of an insufficient number of items or if the examinee volumes are low.

One method to compensate for differences in subtest difficulties would be to have content experts rate the items within each subtest, such that subtest cut scores could be derived using an appropriate standard setting method (e.g., Nedelsky method). Thus, very difficult subtests would be assigned low cut scores, since the content experts would not expect minimally competent examinees to score well on these subtests. Conversely, very easy subtests would be assigned high cut scores. A problem with this approach is that a standard-setting study would probably have to be conducted for each new test form. Since standard-setting studies are often costly and time-consuming to conduct and analyze, such an approach may not be practical.

The reader is warned against the temptation to equate the subtests separately in an effort to maintain the subtest cut scores established by the content experts for an initial (base) test form. Such an action would defeat the purpose of providing useful subtest results, since the equating of subtests from one test form to another is likely to distort the relationship that usually exists between a well-calibrated test and its component subtests. Because subtest and total-test scores would be equated separately, and not

in concert, raw subtest scores would be adjusted in varying directions depending on the outcomes of the individual equatings. As a consequence, examinees who fail a given test form would probably be unable to identify the content-area weaknesses that contributed most to their failure.

Another method of statistically adjusting for differences in subtest difficulties is to use mean-standard deviation linear scaling to convert all subtest scores to a common scale, the scale of the total test. (If the subtest scores are not linearly related to the total test scores and the candidate volume is large, equi-percentile equating should be considered.) Levels of performance could be defined for all subtests jointly, consisting of 3-5 levels ranging from low to high performance. Below is an example based on a 25-100 scale:

Level of Performance	Scores Assigned to Each Level
You have performed at a LOW level of proficiency in this content area--an EXTENSIVE review of study materials is recommended.	25 - 59
You have performed at a MARGINAL level of proficiency in this content area--a COMPLETE review of study materials is recommended.	60 - 75
You have performed at a HIGH level of proficiency in this content area--a BRIEF REVIEW of study materials should be sufficient.	76 - 100

Once the marginal range has been defined, the other two will follow. The marginal range should be defined as the range of subtest scores minimally competent examinees would be expected to earn. Usually this level is initially defined arbitrarily, guided by experience with previous score distributions, and later adjusted as warranted in light of recent empirical findings. Levels of performance should always be defined to identify both weak and strong performance

levels; this would provide examinees with the option of concentrating on either their strengths or weakness. This option should be available, since some examinees prefer the test preparation strategy of eliminating their weaknesses, while other examinees prefer to build their strengths.

To be useful to failing examinees, the subtest results should be consistent with overall pass/fail status. Listed below are some examples of inconsistent results for failing examinees who took a test comprised of four subtests:

Subtest			
1	2	3	4
LOW	HIGH	HIGH	HIGH
LOW	MARGINAL	HIGH	HIGH
MARGINAL	HIGH	HIGH	HIGH
MARGINAL	MARGINAL	HIGH	HIGH

Any of these results would be confusing, in that they would suggest to failing examinees that they should have passed the total test. As several COPA managers indicated during the interview study, inconsistent subtest results have been the subject of many examinees complaints. This usually happens when a passing score is very low for a given test form. While score equating may significantly lower passing scores when the test form being equated is uncharacteristically difficult, such discrepancies accompanied by

significant shifts in passing rates may indicate problems of a statistical or test developmental nature. More often, these inconsistencies occur when performance levels are defined in terms of raw subtest scores, and then applied to future test forms for which total-test passing scores are derived through equating. Performance levels based on raw scores are more susceptible to random fluctuation than passing scores made comparable from one test form to another through equating.

Another method of improving the interpretability of subtest scores was developed by this author for use with a COPA certification test for which the subtest lengths and the examinee volumes were sufficient. The procedure employs the concept of the borderline examinee defined by the standard-setting methods of Angoff (1971) and Nedelsky (1954), in that passing scores are viewed as the average total-test scores borderline or minimally competent examinees would be expected to achieve. Borderline examinees are defined empirically as examinees who score within a defined range centered on the total-test passing score. This group's subtest means are used to derive scaled cut-scores for the subtests, adjusted for differences in difficulty, as well as for consistency with total test pass/fail status. Examinees are provided with a description of three levels of performance, their scaled subtest scores and the ranges of scores assigned to each level. This procedure has met with some success, in that the number of examinee inquiries about interpretation of subtest results has decreased in this program.

Another condition that can impede a failing examinee's effort to identify content areas on which they should concentrate their remedial work is a marked discrepancy among subtest lengths. For example, a 100-item test may consist of an 83-item Subtest A, a 5-item Subtest B and a 12-item Subtest C. Suppose the passing score for the total test is 80% (as a percent-correct score), and that subtest performance levels are defined in terms of the following percent-correct score ranges: "Highly Competent" = above 90%, "Marginally Competent" = 70%-90% and "Deficient" = below 70%. The subtest weighting can lead to misleading score interpretations to failing examinees whose Subtest A performance is Marginally Competent or higher. For example an examinee who fails the total test with a score of 71% (missing the passing score by 9 items) could earn subtest results of:

Subtest A = MARGINAL (75% or 62/83 items correct)

Subtest B = DEFICIENT (40% or 2/5 items correct)

Subtest C = DEFICIENT (58% or 7/12 items correct)

The obvious strategy for the failing examinee is to concentrate his studies on content areas measured by Subtests B and C. However, this would be misleading since, even if the examinee were to retake the same test form and answer 100% of the items in Subtests B and C correctly (while maintaining the same Subtest A score), he would still fail the total test. Perfect scores on these two subtests would only increase his total test score to 79%; the examinee would

fail by one item. The examinee would be better advised to concentrate his efforts on Subtest A, where greater benefits can be expected. In fact, if he were to let his Subtest B and C performance remain the same (Deficient) and merely increase his Subtest A score from 75% to 86% (still within the Marginal range), he would pass the test.

These problems can be minimized at the test planning stage by avoiding large discrepancies among subtest lengths. The process of deciding what proportion of items are assigned to each content area should not only address considerations such the as relative importance of content areas and minimum levels of reliability, but should also address the score interpretive needs of failing examinees.

Another condition that can threaten a failing examinee's interpretation of subtest results is the incomplete reporting of subtest results. That is, when some of the test items in the total test are not assigned to reported subtests. This usually occurs when test content is not adequately defined, or when subtests are not sufficiently lengthy for reporting purposes. For example, a 75-item test may consist of a 20-item subtest and a 30-item subtest, with the remaining 25 items being unassigned to any subtest. Examinee Mary Jones fails this test with a total number-right score of 40 (the passing score was 55); her subtest scores were, respectively, 14 out of 20 (ACCEPTABLE) and 21 out of 30 (ACCEPTABLE). This is a misleading situation for the examinee, since the knowledge that her performance on the subtests was acceptable cannot be very useful to her in improving her low total

test score. It would be far more useful for her to know what content is measured by the unclassified items, since she only answered 5 of those 25 questions correctly.

A final consideration for auditors to weigh in judging whether failing examinees can complete Task 3 is the stability of subtest scores from one form to another. The rationale is that subtest results based on test form A are reported to failing examinees to help them prepare to take test form B. This rationale implies that subtest to subtest reliability is sufficiently high for this purpose. Measuring the stability and equivalence of subtest scores is very difficult, since so many factors operate on examinees during the period between the first testing and subsequent testings. For example, study strategies may be as varied as the number of examinees, especially since testing program administrators tend to have little involvement with examinee training.

A more measurable and relevant study of stability would be to address the extent to which subtest scores decrease on subsequent testings, with an emphasis on high subtest scores. Regression to the mean of a high subtest score on the initial testing may induce examinees to neglect their review of the content area for the second testing. A statistical analysis of the number and extent to which subtest scores of repeat examinees decrease from one testing to a second testing could reveal some information about the practical stability of the subscores. When stability of this type is insufficient, failing examinees may be misled into believing that successful performance on subtest 1 on the first testing implies successful performance on subtest 1 on subsequent testings, given a

brief review of study materials. This is a reasonable assumption for examinees to make; it is the responsibility of measurement statisticians and test developers to ensure that it is also a correct assumption.

IX. DESCRIPTIONS OF EXEMPLARY SCORE INTERPRETIVE INFORMATION

Criteria for identifying score interpretive information (SII) that is minimally acceptable, according to the author, are discussed in the RECOMMENDATIONS section. In keeping with ETS's commitment to quality, efforts to improve the quality of SII, regardless of audit rating, must be ongoing-- development of a minimally acceptable SII should never be the stopping point. The ultimate goal must be the ideal. Described below are two hypothetical scenarios for SII that outline possible directions ETS testing programs could select in improving the quality of their present products. The first scenerio describes a numerically oriented SII, while the second describes a verbally oriented SII.

Scenario 1

Description of Test

In this example, 1,000 candidates for certification in a medical profession are tested once a year. The test is multiple-choice, 200 items in length, consists of four subtests, and is scored as "rights only" (the sum of correct responses).

Some Test Development Steps to Facilitate Score Reporting

Before items were written, test specifications were formulated from the results of a job analysis. The four job dimensions identified by the committee of job incumbents served as content areas to be measured by four subtests. Each subtest was broken down by the job incumbents into subcontent areas for more detailed description. As a result of a joint effort of job incumbents and test developers, the lengths of the subtests were designated, according to the relative importance of the subtests. Each subtest consisted of at least 15 items, so the likelihood of reporting reliable subtest scores would be reasonable. The job incumbents also provided a list of manuals and texts that served as the central reference source for item writers.

Items were written at the content subarea level. At the beginning of the item-writing workshops, the item writers were given a detailed description of the content, a reference list and specifications. Item writers were instructed that all sources for correct item responses must be on the reference list; this condition was verified with the aid of content experts. After an item passed final editing, the references (with page numbers) actually used to develop the item were listed. Final reference lists were prepared at the content area and subarea levels.

Descriptions of Content

Definitions of content areas and subareas were presented in great detail in the candidates' bulletin of information. To explain the test content further, a sample item with answer key was provided for each content subarea.

Development of an Instructional Package

A brief make-up course was developed for each content area in the form of a programmed text. The content and the reference sources of the programmed courses were the same as those used to develop the tests. The programmed texts and the certification tests were thus assumed to be compatible. The programmed texts were experimentally tested, and improvements in test scores were found to be larger for candidates who completed the course than for those who did not. At a reasonable cost, candidates can order a programmed text for each content area they select. This is a potentially lucrative market for ETS licensure and certification testing programs; outside training/coaching programs may teach content that is not measured by the ETS test.

Establishment of Passing Scores

The passing score for the total test was established by a panel of 40 qualified content experts using the Nedelsky method. Following an orientation session to aquaint the judges with the

method and to agree on a definition of the "minimally competent" candidate, the judges were assigned randomly to four groups of 10 to rate the test items. The four studies were conducted concurrently by four well-trained, experienced supervisors. These assignments served to minimize the chances of one or two persons dominating the entire group, as well as to make it possible to identify any serious problems with individual supervisors. The reliability; i.e., intraclass correlation among ratings (Shrout and Fleiss, 1979), was high, .96, with a standard error of judges of less than one.

This client has unlimited financial resources and considers high standards as established by content experts to be far more important than passing rates; the passing score recommended by the judges was selected by the client for score reporting. The subtest passing scores were calculated by averaging the individual judges' ratings tallied by subtest, rather than by total test. The subtest passing scores, or average subtest ratings, were consistent with the total passing score--the total passing score equals the sum of the subtest passing scores, within rounding error. Passing scores of every new test form are established in this manner; equating is not performed.

Reporting Total Test Results

Total test and subtest scores are converted to a scale ranging from 20 to 99; raw passing scores are converted to a scaled score of 70. Scaled total test scores are reported to failing candidates only, to minimize invidious comparisons of scores by candidates and

clients. (Passing candidates receive simply the word "PASS" on their score reports.) Along with their total scaled scores, failing candidates are informed if they scored slightly below, moderately below or well below the passing score. Error measures such as the standard error of measurement at the passing score (Lord, 1984) are used in defining initial score ranges; final selections are based on the results of an analysis to identify the ranges for which classification errors (of assigning candidates to score ranges) are minimal.

Reporting Subtest Results

Subtest scaled scores are reported to failing candidates only. In view of what has been learned from content experts, candidate surveys and statistical analyses, the following three levels were established for all subtest scores:

Level of Performance	Subtest Scores Assigned to Each Level
You have performed at a LOW level of proficiency in this content area--an EXTENSIVE review of study materials is recommended.	20 - 60
You have performed at a MARGINAL level of proficiency in this content area--a COMPLETE review of study materials is recommended.	61 - 80
You have performed at a HIGH level of proficiency in this content area--a BRIEF REVIEW of study materials should be sufficient.	81 - 99

Prescriptive Reading Lists for Failing Candidates

The same reference sources used to build the test and to develop the programmed texts were used to generate prescriptive reading lists for failing candidates. Along with their score reports, failing candidates receive a list of references selected especially for them, on the basis of their item scores. For each incorrectly answered item, one reference is printed under the corresponding content area. For each content area, one reference is printed per line. To facilitate brevity, only nonduplicated references are listed; if the same reference is cited for two or more items, the reference will be listed only once. The references are arranged logically by content subarea. Content areas and subareas are defined both in the candidate bulletin of information and on an insert accompanying score reports. Thus, while content subarea performance would not be reliable enough to report as a score, failing candidates can make good use of the references cited for each content subarea.

An example is provided below for hypothetical candidate John Smith. Mr. Smith failed the total test with a score of 68; his subtest scores were 37(LOW), 70(MARGINAL), 53(LOW) and 81(HIGH) for subtests 1, 2, 3 and 4, respectively. This reading list was selected by a computer program on the basis of his item responses within each subarea of each content area. Findings from a controlled experiment strongly suggest that failing candidates who complete their prescribed readings are significantly more likely to pass a retest than failing candidates who do not.

Prescriptive Reading List for: John Smith

CONTENT AREA

1	2	3	4
(Abbrev. Title)	(Abbrev. Title)	(Abbrev. Title)	(Abbrev. Title)
.....	.....	.....	.....
Subarea 1A	Subarea 2D	Subarea 3B	Subarea 4C
I,pp.26-27	III,pp.12-17	I,pp.5-7	III,p.196
	III,p.20	I,pp.10-13	
		II,pp.2-3	
Subarea 1C			
II,pp.49-52		Subarea 3C	
II,pp.54-59		II,p.4	
II,pp.60-61		II,pp.5-8	
Subarea 1D			
I,p.24			
I,pp.28-32			
II,pp.62-63			
II,p.65			
II,pp.73-76			

.....

Reference Codes:

I = Reference Source X

II = Reference Source Y

III = Reference Source Z

Scenario 2

Description of Test

In this example, 1,500 candidates for certification in a mechanical field are tested twice a year. The test is multiple-choice, 175 items in length, consists of six subtests, and is scored as "rights only."

A Matter of Literacy

Over a period of several years, it became evident to the program director that candidates tested by this program have inadequate reading skills. In addition, candidates repeatedly expressed their confusion and mistrust of numerical scores, urging that their test results be explained in "simple English." Efforts to alleviate these concerns by providing lengthier and more detailed explanations of scaling and equating only worsened the situation. The decision was made to report test results as brief, verbal statements.

Some Test Development Steps to Facilitate Score Reporting

From the job analysis, it was learned that successful performance of entry-level jobs required a sixth-grade reading level, on the average. The test was written at that level and no higher; this was also true for all score interpretive information.

The six subtests were derived from the job dimensions from the job analysis; each subtest measured a class of job skills. The job skills were described in the form of behavioral objectives. To enhance the relationship between the test items and on-the-job performance, previously validated performance tests were identified for each job skill measured by the test. Moreover, the qualified, experienced persons who served as item writers had all the tools and devices necessary to perform the job skills as visual aids.

Concurrent Validity

A controlled experiment was conducted to compare written test and performance test scores. Every one of the job skills measured by the written test related strongly to the corresponding skill measured by the performance tests. The passing score established for the written test was used to classify the candidates into competent and noncompetent groups; these classifications were compared to those defined by the preestablished passing score of the performance test. Written test classifications (i.e., competent or noncompetent) and performance test classifications were the same for 92% of the candidates in the analysis sample. This supportive validity evidence was crucial for instruction and remediation.

Establishment of Passing Scores

The total test passing score was established initially on the basis of a passing score study, using the Angoff method. The passing score recommended by the judges was adjusted by subtracting one standard error of judges, to reduce the probability of failing

truly competent candidates who are disadvantaged by written test items. The passing standard for the total test is maintained through score equating (Tucker and Levine methods).

Passing scores for the subtests were established empirically using the method developed by this author (described in the recommendations section). In brief, candidates scoring within a specified distance of the total test passing score (derived through equating) served as an empirically defined borderline candidate group. (According to the Angoff method, passing scores are derived on the basis of judges ratings of how borderline candidates would be expected to perform on test items.) The empirically defined borderline candidates' mean subtest scores, adjusted for measurement error, serve as passing scores for the subtests. In effect, the borderline candidate group becomes a norm group against which candidates can evaluate their own performance on the subtests.

Test-Taking Aids

Although concurrent validity support of the written test is strong, the client recognizes that some candidates who can successfully perform the job skills are ineffective at taking written tests. A booklet, Test-Taking Aids, was written especially for these candidates; it can be requested at a small charge to first-time candidates and free of charge to repeat candidates. Results from a controlled experiment support the contention that unskilled test-takers can improve their test-taking skills by reading the booklet and completing the exercises.

Sample Tests

A sample test is provided to all candidates in the bulletin of information. This test not only serves to familiarize candidates with the items, but also to improve their understanding of the content area definitions. The 50-item sample test is a scaled-down version of the total tests, representative both in content and statistical specifications. Total sample test passing scores were established by the same procedures used to established the total test passing scores--Angoff derived passing score adjusted by one standard error of judges. In addition to the sample test, the bulletin of information includes the answer key and a list of the items assigned to each of the six content areas.

Reporting Total Test Results

Scaled total test scores were defined as a scale ranging from 25 to 150, with a passing score of 100. The scale is maintained through equating. No scores are reported. Passing candidates are notified that they PASSED, while failing candidates receive one of the following results, ranging in degree of deficiency from "slightly below standard" to "very weak" :

FAILED: Overall performance is slightly below standard.

FAILED: Overall performance is weak.

FAILED: Overall performance is very weak.

These score levels were defined initially for the base test form by

a panel of qualified content experts, and are maintained through equating.

Reporting Subtest Results

A panel of 15 qualified, experienced training course instructors was formed to determine the type of diagnostic information failing candidates should receive. Five levels of performance were designated for reporting subtest results, since most candidates were accustomed to being graded that way in public school. The levels were:

Excellent knowledge development

Good knowledge development

Acceptable knowledge development

Weak knowledge development

Very weak knowledge development

To provide further guidance to failing candidates, the panel developed a remedial self-study guide for each content area. A complete set of these self-study guides has since been routinely mailed to all failing candidates, along with their score reports. Each self-study guide consisted of readings and written and practical exercises.

The panel next identified a remedial prescription for failing candidates at each proficiency level. For example, candidates who have demonstrated "Very weak knowledge development" for content area 1 are advised to complete the self-study guide for that area twice; while candidates who have demonstrated "Excellent knowledge development" for content area 1 are advised merely to read all section summaries. The precise prescriptions differ from one content area to the next, depending on factors such as the difficulty and complexity of the required knowledge.

Content Area	Level of Performance	Recommended Study Assignment
1	Excellent knowledge development	- Read all section summaries.
2	Good knowledge development	- Read all section summaries. - Complete all exercises.
3	Very weak knowledge development	- Complete Study Guide three times.
4	Good knowledge development development	- Read all section summaries. - Complete all practical exercises.
5	Acceptable knowledge development	- Read all section summaries. - Complete all written exercises. - Complete all practical exercises.
6	Weak knowledge development	- Complete Study Guide once. - Read all section summaries.

Sample subtest results appear above for candidate Egbert Adams whose overall performance was "slightly below standard." More detailed study assignments can be developed for testing programs for which content subareas are defined. A research study of failing candidates supported the effectiveness of the remedial prescriptions; failing candidates who indicated that they had performed assignments scored significantly higher on the retest than failing candidates who indicated that they had not. A finding important to the interpretation of the reporting system is that retest scores were positively correlated to the number of times candidates indicated that they completed assignments. For example, candidates who completed their assignment three times tended to score higher than those who completed their assignment twice.

The passing score; i.e., the mean subtest score of borderline candidates, is the lowest subscore at the "Acceptable knowledge development" level. Once the lower limit of acceptable knowledge development is determined for a new test form, the score ranges that define each performance level are set by following the guidelines established by the panel of experts.

X. PROVIDING USEFUL SCORE INTERPRETIVE INFORMATION TO FAILING EXAMINEES: LIST OF RECOMMENDATIONS FOR PROGRAM DIRECTORS, TEST DEVELOPERS AND MEASUREMENT STATISTICIANS

Test Planning

1. Report subtest results only to failing examinees. The reporting of subtest results to passing examinees could prompt examinees and others (e.g., employers) to rank examinees on the basis of their subtest results. Criterion-referenced tests are generally not designed for this purpose.

2. Report total test scaled scores (to failing examinees), not just pass/fail status. Failing examinees should have an indication of how close they came to passing the test.

3. Advise clients that test length limits the number of subtests scores that can be reported. For example, a 60-item test can only yield a maximum of 3-4 subtest scores (subtests at least 15 items in length are required to achieve the desired level of reliability).

4. Advise clients that ALL test items should be assigned to a subtest, and that results for ALL subtests should be reported. No items should be without a subtest assignment, except for tests that do no consist of subtests.

5. Study the content of existing training resources (courses, study guides, etc.) to determine the extent to which they match the content of the test, and are otherwise effective (e.g., predictability of ETS test performance from course grades). ETS and client organizations should be in a position to advise inquiring examinees about the availability of training resources.

Test Specifications

6. Subtests should consist of at least 15 items. This is only a general guideline; if item quality is poor (e.g., inadequately discriminating items) or if some other psychometric problem is identified, the minimum number of items should be higher.

7. Write clearly defined content-area descriptions. Descriptions should be accompanied by well written sample items that delineate the content area.

8. The lengths of subtests should not differ widely. Subtests that are inordinately weighted (e.g., a 60-item subtest in an 80-item test) can cause examinees to misinterpret their subtest results.

9. Reduce content areas to smaller components, content subareas, to be referred to in further clarifying content descriptions for examinees.

Test Development

10. Base item content on written information that is accessible to examinees, in as far as this is possible, given that some necessary test content may only be learned through "on the job" experience.

11. Prepare a reference list of information sources used to write each item. In cases in which printed documentation is not appropriate (e.g., content is learned through "on the job" experience), a written statement of rationale for the keyed response should be prepared. Recurring references may be listed in bulletins of information by content area and subarea to assist failing candidates in preparing for retesting.

Score Interpretive Information

12. Describe content areas clearly and concisely in bulletins of information; whenever possible repeat the descriptions in other documents examinees are expected to refer to in interpreting their score reports (e.g., score report inserts).

13. Provide all examinees with a short sample test assembled according to the specifications of an operational test. The sample test can be distributed with the bulletin of information or with other materials routinely provided to examinees prior to the examination date; it should be accompanied by a scoring key, a list of content-area assignments and a total test passing score.

14. Compile a list of questions and comments failing and passing examinees most frequently ask about interpretation of their score reports. Use the list as a resource for revising score interpretive information (bulletins of information, score reports, etc.).

Scoring of Subtests

15. Define cutoff scores or score ranges for all subtests in order to guide failing examinees in determining their level of performance in each content area measured by the subtests.

16. Set subtest results to be consistent with examinees' total test pass/fail status. Following are examples of consistent and inconsistent test results:

Type of Result	P/F Status	Subtest 1	Subtest 2	Subtest 3
CONSISTENT	FAIL	DEFICIENT	DEFICIENT	GOOD
INCONSISTENT	FAIL	EXCELLENT	GOOD	GOOD

17. If subtest means are not comparable in difficulty, statistical adjustments should be made to enable failing examinees to identify correctly content areas in which they are strong or weak.

.....

APPENDICES

.....

APPENDIX A

List of Testing Programs and Managers Included in Score Reporting Survey and Interview Study

COPA Testing Program in Sample	Manager Interviewed
o American Production amd Inventory Control Society Certification Examination (APICS)	Ms. Melinda A. Peabody Program Director COPA
o National Certification Program for Construction Code Inspectors Examination (CCIP)	" "
o Academy of Certified Social Workers Examination (ACSW)	Mr. Rufus B. Smith, Jr. Program Director COPA
o National Board of Podiatry Examiners Examination (NBPE)	" "
o National Board for Respiratory Care Examinations (NBRC)	" "
o American Institute of Certified Planners Examination (AICP)	Mr. Charles J. Teryek Program Director COPA
o American Journal of Nursing Continuing Education Program Examination (AJN)	" "
o Contact Lens Registry Examination (CLRE)	" "
o International Council of Shopping Centers Examination (ICSC)	" "
o National Association of Housing and Redevelopment Officials Certification Examination (NAHRO)	" "
o National Business Forms Certified Forms Consultant Examinations (NBFA)	" "
o American Board of Opticianry's National Opticianry Competency Examination (NOCE)	" "

o Multi-state Insurance Licensing Program (MILP)	Ms. Patricia L. Testa Assistant Director Custom Operations Div.
o Real Estate Licensing Examination Program (RELE)	" "
o American Society of Heating, Refrigeration and A/C Engineers Examination (ASHRAE)	Ms. Carolyn Y. Uhrich Program Director COPA
o American Society of Plumbing Engineers Certification Examination (CIPE)	" "
o Cosmetology Multi-state Licensing Program (CMLP)	" "
o Energy Auditor Certification Program (EAC)	" "
o Certified Information System Auditor Program (CISA)	Mr. Leo F. Walrad Program Director COPA
o Foreign Service Examination	" "
o Institute of International Container Lessors Examination (IICL)	" "
o Multi-state Electrical Licensing Test (MELT)	" "
o National Association of Purchasing Management Certification Examinations (NAPM)	" "

APPENDIX B

Memorandum Sent to Managers Prior to Their Participation in the Interview Study

MEMORANDUM FOR: MS. DAVIS
MS. PEABODY
MR. SMITH
MR. TERYEK
MS. TESTA
MS. UHRICH
MR. WALRAD

Subject: Interpretation of Failing Test Scores Project: Interviews with Program and Operations Directors

Date: May 2, 1984

From: Gregory A. Shannon

Currently, I am conducting a study to identify information considered to be useful to failing examinees in interpreting score reports. The project is being funded by the Office of Corporate Quality Assurance. As a part of this project, program and operations directors have been asked to talk about score information provided to failing candidates by their programs.

A list of programs to be the focus of your particular interview is attached. These programs are among 23 testing programs selected to comprise a representative sample of COPA testing programs that report test results to failing candidates.

Also attached is a list of the interview questions you will be asked. Your interview will be about a half-hour in duration, and will be conducted by me at your convenience. You can charge your time to P/J: 01609.

Thank you for your cooperation.

APPENDIX C

IFTS Project: Managers Interview Questions

INSTRUCTIONS TO INTERVIEWEE

"Please answer the following questions only in reference to your programs listed on this list. (Provide interviewee with list of his/her programs in IFTS sample.)

"Before we begin, I should explain what I mean by the terms: 'survey' and 'candidate score information.'

"The term 'survey' refers to the systematic collection of facts, opinions and comments from people sampled from a target population of interest (e.g., all people registered to vote in New Jersey in November 1984), through questionnaire mailings, telephone interviews or other methods.

"Candidate score information is information sent to candidates for the purpose of reporting or explaining their test results. This would include information such as score reports, descriptive inserts enclosed with score reports or 'score interpretation' sections of bulletins of information."

1. Has anyone ever surveyed candidates who failed examinations administered by any of your programs to solicit their evaluations of the usefulness and clarity of candidate score information sent to them?

   Yes ----- No -----

   If yes, which programs?

2. Please answer the following questions for each program for which a survey of this type was conducted. (Call the name of each program listed above.)

   ..... What survey method was used (e.g., telephone interviews, questionnaire mailings, etc.)?

   ..... What is the name of the person and organization who conducted the survey?

   ..... When was it conducted?

   ..... How many candidates responded?

   ..... What was the approximate total number of candidates targeted by the survey?

   ..... Please specify the major findings.

..... Were any of the findings useful in revising candidate score information?

Yes ----- No -----

..... If yes, which findings were useful?

..... If yes, which changes resulted?

3. Do you consider a survey of this type to be useful for any of your programs?

Yes ----- No -----

If yes, which programs?

..... How might your programs benefit from this type of survey?

..... Do you think all of your clients, of the programs listed, would consent to a survey of this type being conducted?

Yes ----- No -----

If no, which clients would not consent?

4. The following questions pertain to the score reports that are currently being issued to failing candidates by your programs. Please answer each question separately for each program. (For each question, call the name of each program on the list.)

..... When were your score reports first developed?

..... How frequently are score reports updated?

..... What influences your decisions to update or modify score reports for your program? (Say to interviewee: "Influences might include client requests, candidate comments, COPA staff recommendations and so forth.")

5. Do you think candidate score information currently provided by any of your programs could be improved, to be more useful to failing candidates?

Yes ----- No -----

If yes, which programs?

..... How might your candidate score information be improved to be more useful to failing candidates? (Call the name of each program listed above.)

..... Assuming all pertinent improvements could be made, how might your client benefit? (Call the name of each program listed above.)

6. Based upon your experience with all of your programs (on interviewer's list), what questions and comments are most frequently asked by failing candidates, concerning the interpretation of their score reports?

7. Please answer each of the following questions separately for each program. (For each question, call the name of each program on the list.) Which person or organization would be the first to respond to questions and comments from failing candidates, concerning the interpretation of score reports?

..... What procedure would be followed if this person or organization cannot satisfactorily respond to failing candidates' questions and concerns about their score reports?

..... Of the total time your staff spends working on one test administration, what percentage of this time is spent responding to failing candidates' questions and comments about their score reports?

BIBLIOGRAPHY

Ahmann, J.S. Basic Issues Concerning Competency-Based Testing. In R.B. Ingle, M.R. Carroll, and W.J. Gephart (Eds.), THE ASSESSMENT OF STUDENT COMPETENCE IN THE PUBLIC SCHOOLS. Bloomington, Indiana: Phi Delta Kappa, 1978, 78-89.

Angoff, W.H. Scales, Norms, and Equivalent Scores. In R.L. Thorndike (Ed.), EDUCATIONAL MEASUREMENT (2nd ed.). Washington, District of Columbia: American Council on Education, 1971, 514-515.

EDUCATIONAL TESTING SERVICE STANDARDS FOR QUALITY AND FAIRNESS. Princeton, New Jersey: Educational Testing Service, 1983.

EDUCATIONAL TESTING SERVICE STANDARDS FOR QUALITY AND FAIRNESS. Princeton, New Jersey: Educational Testing Service, 1981.

Furlong, F. and Miller, W. Diagnose: Computer-Based Reporting of Criterion-Referenced Test Results. EDUCATIONAL TECHNOLOGY. 1978, 18, 37-39.

Glaser, R. and Nitko, A.J. Measurement in Learning and Instruction. In R.L. Thorndike (Ed.), EDUCATIONAL MEASUREMENT (2nd ed.). Washington, District of Columbia: American Council on Education, 1971, 625-670.

Gropper, G.L. DIAGNOSIS AND REVISION IN THE DEVELOPMENT OF INSTRUCTIONAL MATERIALS. Englewood Cliffs, New Jersey: Educational Technology Publications, 1975.

Haladyna, T. and Roid, G. A COMPARISON OF DECISION-MAKING METHODS FOR CRITERION-REFERENCED TESTS. Paper presented at the annual meeting of the American Educational Research Association, Boston, Massachusetts, April 1980.

Hambleton, R.K. and Eignor, D.R. Unit 8: Using and Reporting Test Score Information. A PRACTIONER'S GUIDE TO CRITERION-REFERENCED TEST DEVELOPMENT, VALIDATION, AND TEST SCORE USUAGE, Laboratory of Psychometric and Evaluative Research Report No. 70. Amherst, Massachusetts: School of Education, University of Massachusetts, 1979.

Hambleton, R.K., Gorth, W.P., and O'Reilly, R.P. An Application of an Evaluation Model for Classroom Instruction. JOURNAL OF EDUCATIONAL TECHNOLOGY SYSTEMS, 2(2), 1973, 117-131.

Lord, F.M. Standard Errors of Measurement at Different Ability Levels. JOURNAL OF EDUCATIONAL MEASUREMENT, 21(3), 1984, 239-243.

McArthur, D.L. and Choppin, B.H. Computerized Diagnostic Testing. JOURNAL OF EDUCATIONAL MEASUREMENT, 21(4), 1984, 391-397.

Millman, J. STRATEGIES FOR CONSTRUCTING CRITERION-REFERENCED ASSESSMENT INSTRUMENTS. Paper presented at the Eighth Annual Conference on Large-Scale Assessment, Boulder, Colorado, 1978.

Mills, C.N. and Hambleton, R.K. GUIDELINES FOR REPORTING CRITERION-REFERENCED TEST SCORE INFORMATION. Paper presented at the annual meeting of the American Educational Research Association, Boston, Massachusetts, April 1980.

Nedelsky, L. Absolute grading standards for objective tests. EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 1954, 14(1), 3-19.

Nitko, A.J. Distinguishing the Many Varieties of Criterion-Referenced Tests. REVIEW OF EDUCATIONAL RESEARCH, 1980, 50(3), 461-485.

Payne, D.A. THE ASSESSMENT OF LEARNING: COGNITIVE AND AFFECTIVE. Lexington, Massachusetts: D.C. Heath and Co., 1974, 334-336.

Popham, W.J. Domain Specifications Strategies. In R.A. Berk (Ed.), CRITERION-REFERENCED MEASUREMENT: THE STATE OF THE ART. Baltimore, Md.: The Johns Hopkins University Press, 1980, 15-31.

Popham, W.J. The Case for Criterion-Referenced Measurements. EDUCATIONAL RESEARCHER, 7, 1978a, 6-10.

Popham, W.J. CRITERION-REFERENCED MEASUREMENT. Englewood Cliffs, N.J.: Prentice-Hall, 1978b.

Shrout, P.E. and Fleiss, J.L. Intraclass Correlations: Uses in Assessing Rater Reliability. PSYCHOLOGICAL BULLETIN, 86 (2), 1979, 420-428.

Simpson, M. and Arnold, B. Diagnostic Tests and Criterion-Referenced Assessments: Their Contribution to the Resolution of Pupil Learning Difficulties. PROGRAMMED LEARNING AND EDUCATIONAL TECHNOLOGY, 20(1), 1983, 36-42.

Spencer, E.F. and Smith, R.M. Diagnosing Arithmetic Skills. In R.M. Smith (Ed.), TEACHER DIAGNOSIS OF EDUCATIONAL DIFFICULTIES. Columbus, Ohio: Charles E. Merrill Publishing Co., 1969, 163-164.

Wilson, S.M. and Hiscox, M.D. Using Standardized Tests for Assessing Local Learning Objectives. EDUCATIONAL MEASUREMENT: ISSUES AND PRACTICE, 3(3), 1984, 19-22.